基于深度学习的中文机构名识别研究*
——一种汉字级别的循环神经网络方法

朱丹浩[1,2]　杨　蕾[3]　王东波[4]
[1](江苏警官学院图书馆　南京 210031)
[2](南京大学计算机科学与技术系　南京 210093)
[3](南京交通技师学院中(高)职教育处　南京 210049)
[4](南京农业大学信息科学技术学院　南京 210095)

摘要: 【**目的**】中文机构名结构复杂、罕见词多, 识别难度大, 对其进行正确识别对于信息抽取、信息检索、知识挖掘和机构科研评价等情报学中的后续任务意义重大。【**方法**】基于深度学习的循环神经网络(Recurrent Neural Network, RNN)方法, 面向中文汉字和词的特点, 重新定义了机构名标注的输入和输出, 提出汉字级别的循环网络标注模型。【**结果**】以词级别的循环神经网络方法为基准, 本文提出的字级别模型在中文机构名识别的准确率、召回率和 F 值均有明显提高, 其中 F 值提高了 1.54%。在包含罕见词时提高更为明显, F 值提高了 11.05%。【**局限**】在解码时直接使用了贪心策略, 易于陷入局部最优, 如果使用条件随机场算法进行建模可能获取全局最优结果。【**结论**】本文方法构架简单, 能利用到汉字级别的特征来进行建模, 比只使用词特征取得了更好的结果。

关键词: 机构名识别　循环神经网络　深度学习
分类号: G351

1 引　言

机构泛指机关、团体或其他企事业单位, 包括院校、公私企业、政府部门、宗教组织、科研部门、国际组织、体育团队、音乐团体、军队等[1]。机构名的识别效果对信息抽取、信息检索、知识挖掘和机构科研评价等后续任务起着重要的影响。然而, 中文机构名中罕见词多、结构复杂, 不同机构的名称差异性较大, 这些问题对正确识别机构名带来了很大的挑战。

中文机构名识别可以看做一个序列化标注问题, 基于人工特征模板的模型是解决这一类问题的主要手段, 使用的算法包括条件随机场[2]、支持向量机[3]和最大熵模型[4]。这一类方法面向中文机构名内部和外部的特征, 人工设计了区分性强的特征模板, 然后使用一个强大的序列化标注模型进行标注, 取得了较好的识别效果。但是, 此类方法依赖于专家的领域知识, 在不同类型的语料上难以移植和泛化。近年来, 通过深度学习的策略, 基于循环神经网络的方法在英文的序列化标注领域取得了较大的成功, 包括词性标注、汉语分词、组块分析、命名实体识别和语义角色标注等任务[5-6]。循环神经网络不特别需要人工制定规则, 可以自行从分布式词向量中学习出特征以供标注使用, 逐渐成为研究的热点。

循环神经网络的主要输入是词向量, 词向量的质量直接决定了系统的性能。对于罕见词, 模型不能获取足够的上下文信息, 因此学习出的词向量质量很

通讯作者: 朱丹浩, ORCID: 0000-0003-0477-8517, E-mail: jisuanyuyan@163.com。

*本文系江苏省高校哲学社会科学项目"高校危机管理案例知识库构建及知识挖掘研究"(项目编号: 2014SJB246)、江苏省警官学院"公安学术语自动抽取技术研究"(项目编号: 2015SJYZQ01)和国家自然科学基金项目"基于 CSSCI 的句法级汉英平行语料库构建及知识挖掘研究"(项目编号: 71303120)的研究成果之一。

差。有些研究使用复杂的规则，从汉字中获取信息以强化词向量中的信息。Chen 等使用词中的每一个字来加强中文词汇的词向量，为了解决字的歧义性，首先对字进行聚类，对不同类中的字使用不同的字向量[7]。Sun 等使用部首对中文词向量进行加强，在比较字相似度任务和中文分词任务上取得一定提高[8]。

然而，构架简单、易于泛化是循环神经网络的主要优势，这些复杂的词向量增强方法虽然可以一定程度上解决词向量信息稀疏的问题，却由于规则复杂、实现困难，弱化了循环神经网络的优势。针对以上问题，本文提出一种完全基于汉字的中文机构名识别方法，重新定义了模型的输入和输出。输入为汉字和空格，输出为一套新的机构名标记。该方法结构简单、易于实现，不需要添加任何人工规则和外部资源。

本文的贡献主要有两点：将循环神经网络应用到中文机构名识别任务上，验证了使用深度学习进行中文机构名识别的有效性；针对中文字词特点对标注模型进行改进，取得了更好的标注效果。

2　相关研究

作为经典的序列化标注任务，机构名的识别一直是情报学关注的重点之一。近年来，以循环神经网络为主的深度学习方法在序列化识别领域取得了新的进展。本文将从命名实体识别和循环神经网络两个方向对相关研究进行梳理。

2.1　命名实体识别相关研究

命名实体的识别策略主要围绕着基于规则和基于统计两种方法展开，但以统计方法为主，比较有代表性的方法如下。孙镇等从技术方法和评价两个角度对命名实体的研究情况进行了系统而详细的论述[9]。在构建的内部和外部规则基础上，潘正高提出了基于概率统计的命名实体识别策略[10]。陆伟等在条件随机场模型的基础上，完成了对产品命名实体的识别[11]。从跨语言检索的角度，吴丹等给出了翻译加权的命名实体策略[12]。基于条件随机场，王文龙等通过统计项目申请书中的各种命名实体的特征，构建了多特征知识下的命名实体识别模型[13]。结合词性与知网的外部语义特征知识，陈锋等结合条件随机场完成了对学术期刊中理论这一实体的自动识别[14]。俞鸿魁等提出一种基于角色标注的中文机构名自动识别方法，根据在机构名识别中的作用，采取 Viterbi 算法对切分结果进行角色标注，在角色序列的基础上，进行字符串识别，最终实现中文机构名的识别[15]。关晓炟等提出一种自动构建用户查询日志机构名训练语料的方法，解决目前用户查询日志语料资源匮乏的问题[16]。利用粘合度概念解决信息不对称问题，结合上下文等信息，采用条件随机场模型进行机构名识别。基于统计的方法在不同的语料上可以有效识别出机构名，但依赖于专家对具体语料提出的规则和特征模板，方法复杂且难以移植。

2.2　循环神经网络

循环神经网络在许多英文序列化标注任务上表现出强劲的标注能力，在循环神经网络的算法框架下，使用长短期记忆模块(Long Short Term Memory, LSTM)来代替基本的 TANH 模块会取得更好的效果。Huang 等使用双向 LSTM 进行序列化标注，并在输出层使用条件随机场(Conditional Random Fields, CRF)进行解码，在多个数据集上对词性标记、组块分析和命名实体识别任务进行验证，发现在加入人工规则和预训练词向量后，该方法达到了最好性能[17]。Ma 和 Hovy 使用双向 LSTM-CNNS-CRF 模型实现了端对端的序列化标注，使用卷积神经网络(Convolutional Neural Networks, CNNS)对每一个词学习出字级别向量，然后将字级别向量和词向量拼接成一个加强向量，输入到双向 LSTM 模型中，最后使用条件随机场进行解码，在英文词性标注和命名实体两个任务上验证了该方法[18]。虽然在英文方面已经有研究者开始探索在循环神经网络中增加字信息来进行建模，但中文方面尚缺乏类似研究。英文和中文在字和词上有较大的差异，本文针对中文字和词的特点，设计了新的算法来使用汉字信息。

3　系统框架和模型

3.1　系统框架

图 1 给出了标注系统的框架，总共分 4 层。最下面一层是第一层，原始模型的输入为词，本文提出的字模型输入字和分词标记。第二层为向量映射层，将第一层的输入转化为对应的分布式表示向量。第三层为循环神经网络层，图中展示的是一个两层的 LSTM 循环神经网络。最上面一层是输出层，循环神经网络的结果在这一层被转换为输出标记。

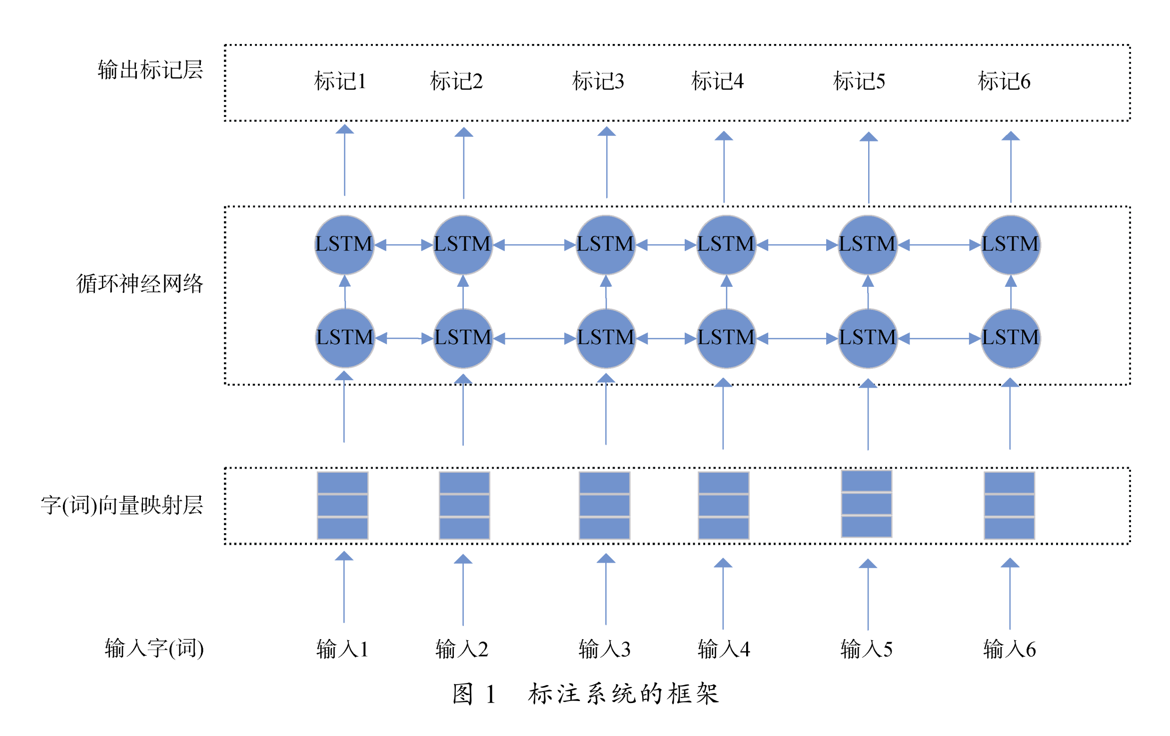


图 1　标注系统的框架

本文改进了第一层输入层和顶层输出层。为了表述简洁，下面直接用"LSTM"代替"LSTM 循环网络"，因为目前 LSTM 节点只能应用于循环网络之中。

3.2　循环神经网络相关模型

循环神经网络(Recurrent Neural Network, RNN)是一种特别适合序列化标注的神经网络模型。在循环神经网络中，在时间 t 时刻输入一个向量 $x_t \in R^n$，结合前一步的隐藏层向量 $h_{t-1} \in R^m$，生成当前的隐藏层状态向量，如公式(1)所示:

$$h_t = f(Wx_t + Uh_{t-1} + b) \tag{1}$$

其中，$W \in R^{m\times n}$，$U \in R^{m\times m}$，$b \in R^m$ 是模型中的系数矩阵，f 是激活函数。最后，可以在隐藏状态层之上加上 Softmax 层来进行分类任务，因此，可以理解成 RNN 的输入是 x，输出是 h。

从理论上来讲 RNN 可以保留住长距离记忆，但在实践中，由于梯度消失和梯度爆炸现象，原始的 RNN 模型难以做到这一点。Hochreiter 等和 Sutskever 等对原始的 RNN 进行改进，提出了长短期记忆模块(LSTM)，通过在 RNN 中增加记忆模块和一些控制阀解决了长距离记忆问题[19-20]。一个标准的 LSTM 模块[20]如下所示:

$$i_t = \sigma(W^i x_t + U^i h_{t-1} + b^i) \tag{2}$$

$$f_t = \sigma(W^f x_t + U^f h_{t-1} + b^f) \tag{3}$$

$$o_t = \sigma(W^o x_t + U^o h_{t-1} + b^o) \tag{4}$$

$$g_t = \tanh(W^g x_t + U^g h_{t-1} + b^g) \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{6}$$

$$h_t = o_t \odot \tanh(c_t) \tag{7}$$

在 t 步，通过前一步的记忆模块 c_{t-1}，前一步的隐藏状态 h_{t-1} 和当前输入 x_t 来计算当前步的隐藏状态 h_t 和当前记忆 c_t。$\sigma(\cdot)$ 和 $\tanh(\cdot)$ 分别是 Sigmoid 函数和正切函数。i_t, f_t, o_t, g_t 分别利用前一状态和当前输入作为控制阀来控制模型的输入输出，以及记忆的转移和保存。由于记忆模块的转移使用了加法运算符，在进行反向梯度计算时解决了矩阵乘法带来的梯度消失和梯度爆炸现象。

在 LSTM 网络中，如果将多个隐藏状态层叠加，低层的输出作为高层的输入，这就形成深层长短期记忆模型(Deep LSTM)。简单的 LSTM 网络是从左向右依次计算的，如果在计算隐藏状态时同时从右向左进行，则称为双向长短期记忆模型(Bi-directional LSTM)。如无特殊说明，下面的基于 LSTM 的序列化标注方法，均使用了 Bi-directional LSTM。

3.3　基于词的机构名标注模型

使用 LSTM 基于词进行机构名标注相当直观，图 2 给出一个机构名标注的示例。最下面一层是输入层，输入层的每个词属于一个有限的集合词汇表 V，中间虚线框内为向量映射层和 LSTM 循环神经网络，这里

同图 1，所以隐藏细节。最上面一层是输出层，输出了对应的标记，标记属于一个有限的集合标记表 S。本文使用三元标记集{B-ORG, I-ORG, S}，B-ORG 表示机构名的第一个词，I-ORG 表示机构名的其余词，S 表示不属于机构名的词。

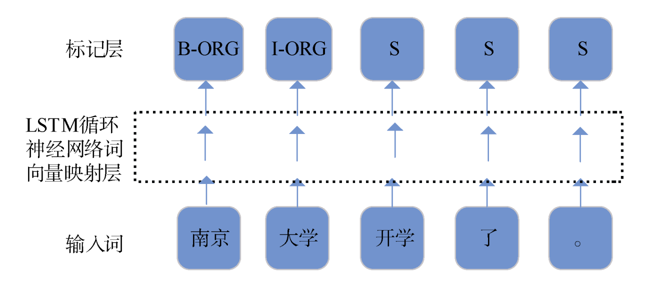


图 2 基于词的机构名标注模型示例

LSTM 在 t 时刻的输入是向量 $x_t \in R^n$，因此要将输入的词 $v_t \in V$ 转换为向量 x_t，x_t 称为 v_t 的词向量。设一个 $k \times n$ 维稠密向量矩阵 L，k 为 V 的词的数量，则 L 中的每一列一一对应于 V 中的词。将输入词 v_t 转换为向量 x_t，只需要根据 v_t 在 V 中的序号到 L 中查找即可。还需要根据隐藏状态 h_t 计算当前的标记 $s_t \in S$ 的概率。这里使用简单的 Softmax 函数，如公式(8)所示。

$$p(s_t = k) = \frac{e^{f(k)}}{\sum_{k' \in s} e^{f(k')}} \quad \text{for } k \in S \tag{8}$$

$f(\cdot)$ 将状态 h_t 线性变换为实数，$f(k) = w_k^T h_t + b_k$，其中 w_k^T 为 n 维系数向量，b_k 为 bias 项。本文使用交叉熵来计算损失函数，时刻 t 的损失函数为公式(9)。

$$J(t) = -y_t(k)\log(p(s_t = k))$$

$y_t(k)=1$ 如果 t 步的真实标记为 k，否则等于 0 (9)

总的损失函数为每一步的损失之和，如公式(10)所示。

$$J = \sum_t J(t) = \sum_t -y_t(k)\log(p(s_t = k)) \tag{10}$$

模型需要学习的参数包括 LSTM 本身的参数，词向量矩阵 L，计算标记概率时的参数 w_k，b_k，k 对应于每一个标记。

3.4 基于汉字的机构名标注模型

图 3 给出了基于汉字的机构名标注示例。在输入层，输入的不再是词，而是一个个的汉字和分词符号<GO>。<GO>表示其下一个输入字符和前一个输入字符不属于同一个词。LSTM 层和词模型没有区别。在输出层，只有<GO>对应的位置才输出标签。

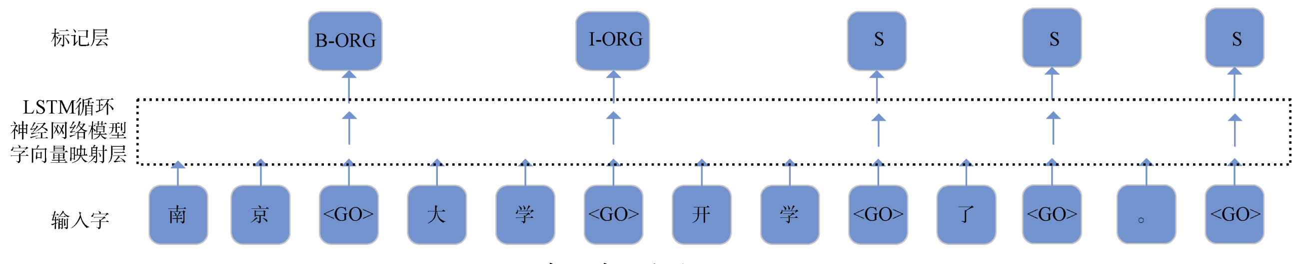


图 3 基于字的机构名标注示例

<GO>这个标记在模型中起着关键性的作用。一开始，笔者设计的模型中并没有<GO>，而是直接在每个词的最后一个字符输出这个词的标签，这样会和现在的做法在标注结果上有一定差距。因为模型中的每一个字都有可能输出标签、也有可能不输出，<GO>实际上起到了告诉模型输出位置的作用。在序列到序列的转换模型中，比如神经网络机器翻译，先按词逐个输入一串源语言句子，在句子的末尾增加一个<EOF>标记表示句子结束，在<EOF>开始输出目标语言的词汇，这里的<EOF>和本文的<GO>标记起着相似的作用。

此模型输入是字，也要先通过一个查找表将字转化为字向量输入 LSTM，输出时只在<GO>对应的位置计算损失函数。因此本文提出新的总损失函数如下：

$$J = \sum_{i(t)=<GO>} J(t) = \sum_{i(t)=<GO>} -y_t(k)\log(p(s_t = k)) \tag{11}$$

其中，i(t)表示第 t 步的输入字。

如果按照测试集的输入来看，中文机构名识别的实验设定有两种。第一种是输入原始文本，系统构建分词和机构名识别一体化模型，如周俊生等的研究[2]。第二种是在分好词的语料上进行机构名识别，如潘正高的研究[10]。本文的设定按照第二种方式进行。两种设定都有可能用到字特征，以特征模板的方式呈现，但与本文的用法完全不同，本文是直接将字作为基本单元输入。

4 实 验

4.1 数据集和评价指标

基于北京大学计算语言学研究所发布的 1998 年上半年的人民日报语料进行模型性能的测试。人民日报语料已经分好词，标注了机构名，机构名被标注为“nt”。以词模型为基准，验证字模型的表现。在实验中，以 1998 年 2 月份的数据为测试集, 1998 年 1 月、3 月、4 月、5 月和 6 月的数据为训练集。对于词模型，根据训练数据建立的词表大小为 40 000，包括 39 999 个频数较高的词和 1 个罕见词标记“RAREWORD”。在测试集中，凡是在词表中没有出现的词均标记为罕见词。按照相同的方法，笔者建立了字表，字表的大小为 5 500，包括 5 498 个频次较高的字、1 个罕见字标记“RARECHAR”和 1 个分词符号“GO”。对于语料库中的词、字和罕见字词的统计数据如表 1 所示。在测试集中，罕见词占比 3.18%，罕见字只有 0.01%，几乎可以忽略不计，这说明基于字的方法将更少遇见未登录现象。

表 1　人民日报 1998 年上半年语料统计数据

对比项目	训练集	测试集
词数	6 137 295	1 149 581
罕见词	144 515	36 508
字数	10 097 274	1 878 731
罕见字	90	218

评价机构名标注时使用三个指标，分别是准确率 P、召回率 R 和 F 值，计算方法如公式(12)–公式(14)所示。其中 T 表示标注正确的机构数，M 表示测试集中的机构数, N 表示标注出的实体数。

$$P = \frac{T}{M} \tag{12}$$

$$R = \frac{T}{N} \tag{13}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{14}$$

4.2 参数设置

使用小批量随机梯度下降法进行反向梯度传递，设置批量为 20，初始的学习率为 1.0，在第 5 轮迭代时，开始按 0.8 的速度减少学习率，总计学习 13 轮。词模型的反向传递的最大步数为 35，由于字模型的步数要比词模型大，将字反向传递的最大步数设为 55。初始化所有的参数为–0.1 至 0.1 之间的随机分布。为了防止梯度过大，使用梯度夹子(Gradient Clipping)技术并设置为 5.0 [21]。为了减轻过拟合现象，使用 Dropout 技术[22]，并设置为 0.8。

4.3 实验结果

表 2 给出了隐藏层为 2 层，隐藏层维度为 650 时的识别结果。“总体”指模型对于所有机构名识别的性能，“包含罕见词”指机构名中包含一个或者一个以上较罕见词的情形。可以看出，总体上，字模型的准确率较基准要高 1.23%，召回率高 1.82%, F 值高 1.54%。在罕见词上的表现尤为突出，准确率要高 8.87%，召回率超出 12.37%, F 值超出 11.05%。罕见词的指标高，说明本文方法在迁移到不同的语料时，具有巨大的优势。

表 2　字模型和词模型的机构名识别结果

任务	总体			包含罕见词		
	准确率	召回率	F 值	准确率	召回率	F 值
字模型	91.87%	88.65%	90.23%	89.86%	76.96%	82.91%
词模型(基准)	90.64%	86.83%	88.69%	80.99%	64.59%	71.86%

对标注结果进行了一定的分析，在举例时，中括号括起的部分表示机构名。字模型误标机构名时，降低了模型的召回率，主要情况包括两种。第一种是语料库漏标或者有争议性的机构名。例如，“电气化 局 三处 是 [铁道部]首家 通过……”，语料库中漏标，但算法可以正确识别，类似的还有“[国家 森林 管理 局]”, “[曲靖 电厂]”等。有争议性的例如“图 为 新东安 市场 [中安天平 图书 中心] 一角”，但模型中识别为“图 为 [新东安 市场 中安天平 图书 中心]一角”，根据笔者的观点，字模型的结论也有一定的道理，以地址为机构名前缀也并无不可。第二种，将罕见的地名识别为机构名，这是因为地名常常是机构名的一部分，

有时会误判，例如“委内瑞拉”、“瑞士”等。根据对结果的观察，第一种是导致召回率下降的主要因素。

字模型漏标机构名时，降低了模型的准确率，主要有以下几种情况。首先，最多情况是整体标注不合法。一个合法的机构名标注应当以 B-ORG 开头，后面的是 I-ORG，例如“[中 俄 总理 定期 会晤 委员会]”标注为“B-ORG I-ORG I-ORG I-ORG I-ORG I-ORG”，但模型中标注为“B-ORG S I-ORG I-ORG I-ORG I-ORG”。第二种是由训练语料不足导致的，例如“[绵阳 国家级 高新技术 产业 开发区]”，在模型中出现多次识别错误，这是因为“锦阳”在训练语料中出现频次较低，且极少出现在机构名中，所以模型不能正确识别。

综上，召回率的错误主要是因为语料库漏标或者有争议性标注，准确率的错误主要是由于语料不足和标注不合法。针对语料不足问题，未来将使用大规模的无标注语料来训练字向量，同时引入多任务学习等技术来缓解；对于标注不合法问题，可以使用 CRF 模型对于输出进行约束，更为精准地搜索结果。

5 结 语

本文针对汉字和词的特点，基于循环神经网络中的双向深度 LSTM，提出汉字级别的中文机构名标注模型，与基准的词级别模型相比，字模型在识别能力上有明显提高，特别是罕见词上的标注能力要大大超出，这说明本文模型在迁移到新语料时有很大的优势。受益于深度学习的特点，相较于传统的特征模板类方法，本文模型是完全端对端的，不再依赖于人工置顶规则，更为简单易用。在未来的工作中，将进一步探索其他深度学习方法在中文序列化标注上的应用，并将尝试新的方法以提高模型的标注能力。

参考文献：

[1] 沈嘉懿，李芳，徐飞玉，等. 中文组织机构名称与简称的识别[J]. 中文信息学报，2007, 21(6): 17-21. (Shen Jiayi, Li Fang, Xu Feiyu, et al. Recognition of Chinese Organization Names and Abbreviations [J]. Journal of Chinese Information Processing, 2007, 21(6): 17-21.)

[2] 周俊生，戴新宇，尹存燕，等. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报，2006, 34(5): 804-809. (Zhou Junsheng, Dai Xinyu, Yin Cunyan, et al. Automatic Recognition of Chinese Organization Name Based on Cascaded Conditional Random Fields[J]. Acta Electronica Sinica, 2006, 34(5): 804-809.)

[3] 黄德根，李泽中，万如. 基于 SVM 和 CRF 的双层模型中文机构名识别[J]. 大连理工大学学报，2010, 50(5): 782-787. (Huang Degen, Li Zezhong, Wan Ru.Chinese Organization Name Recognition Using Cascaded Model Based on SVM and CRF [J]. Journal of Dalian University of Technology, 2010, 50(5): 782-787.)

[4] 滕青青，吉久明，郑荣廷，等.基于文献的中文命名实体识别算法适用性分析研究[J]. 情报杂志，2010, 29(9): 157-161. (Teng Qingqing, Ji Jiuming, Zheng Yongting, et al. Applicability Analysis of Chinese Named Entity Recognition Method Based on Literatures [J]. Journal of Intelligence, 2010, 29(9): 157-161.)

[5] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging [OL]. arXiv Preprint.arXiv: 1508.01991.

[6] Chen X, Qiu X, Zhu C, et al. Gated Recursive Neural Network for Chinese Word Segmentation [C]. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.2015: 1744-1753.

[7] Chen X, Xu L, Liu Z, et al. Joint Learning of Character and Word Embeddings [C]. In: Proceedings of the 24th International Conference on Artificial Intelligence. 2015: 1236-1242.

[8] Sun Y, Lin L, Yang N, et al. Radical-enhanced Chinese Character Embedding [C]. In: Proceedings of the International Conference on Neural Information Processing. Springer International Publishing, 2014: 279-286.

[9] 孙镇，王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术，2010(6): 42-47. (Sun Zhen, Wang Huilin. Overview on the Advance of the Research on Named Entity Recognition [J]. New Technology of Library and Information Service, 2010(6): 42-47.)

[10] 潘正高. 基于规则和统计相结合的中文命名实体识别研究[J]. 情报科学，2012, 30(5): 708-712. (Pan Zhenggao. Research on the Recognition of Chinese Named Entity Based on Rules and Statistics [J]. Information Science, 2012, 30(5): 708-712.)

[11] 陆伟，鞠源，张晓娟，等. 产品命名实体特征选择与识别研究[J]. 图书情报知识，2012(3): 4-12. (Lu Wei, Ju Yuan, Zhang Xiaojuan, et al. Research on Product Named Entity Feature Selection and Recognition [J]. Document, Information & Knowledge, 2012(3): 4-12.)

[12] 吴丹, 何大庆, 陆伟. 跨语言信息检索中的命名实体识别与翻译[J]. 图书情报知识, 2012(3): 13-19. (Wu Dan, He Daqing, Lu Wei. The Extraction and Translation of Named Entity in Cross Language Information Retrieval [J]. Document, Information & Knowledge, 2012(3): 13-19.)

[13] 王文龙, 王东波. 面向项目申请书的命名实体抽取模型构建研究[J]. 情报资料工作, 2015(1): 30-34. (Wang Wenlong, Wang Dongbo. Project Application-oriented Named Entity Extraction Model Construction [J]. Information and Documentation Services, 2015(1): 30-34.)

[14] 陈锋, 翟羽佳, 王芳. 基于条件随机场的学术期刊中理论的自动识别方法[J]. 图书情报工作, 2016, 60(2): 122-128. (Chen Feng, Zhai Yujia, Wang Fang. Automatic Theory Recognition in Academic Journals Based on CRF [J]. Library and Information Service, 2016, 60(2): 122-128.)

[15] 俞鸿魁, 张华平, 刘群. 基于角色标注的中文机构名识别[C]. 见: 第 20 届东方语言计算机处理国际会议论文集. 2003: 79-87. (Yu Hongkui, Zhang Huaping, Liu Qun. Recognition of Chinese Organization Name Based on Role Tagging [C]. In: Proceedings of the 20th International Conference on Computer Processing of Oriental Languages. 2003: 79-87.)

[16] 关晓炟, 吕学强, 李卓, 等. 用户查询日志中的中文机构名识别[J]. 现代图书情报技术, 2014(1): 72-78. (Guan Xiaoda, Lv Xueqiang, Li Zhuo, et al. Chinese Organization Name Recognition in User Query Log [J]. New Technology of Library and Information Service, 2014(1): 72-78.)

[17] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging [OL]. arXiv: 1508.01991.

[18] Ma X, Hovy E. End-to-End Sequence Labeling via Bi-directional LSTM-CNNs-CRF [OL]. arXiv Preprint. arXiv: 1603.01354.

[19] Hochreiter S, Schmidhuber J. Long Short-term Memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[20] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks [A]. //Advances in Neural Information Processing Systems [M]. 2014: 3104-3112.

[21] Pascanu R, Mikolov T, Bengio Y. On the Difficulty of Training Recurrent Neural Networks [J]. Journal of Machine Learning Research, 2013, 28(3): 1310-1318.

[22] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

作者贡献声明:

朱丹浩: 提出研究思路, 设计和进行实验, 论文撰写;
杨蕾: 协助设计研究方案, 负责数据预处理;
王东波: 论文修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: jisuanyuyan@163.com。

[1] 朱丹浩, 杨蕾, 王东波. data cleaning programming.zip. 基于人名日报的语料预处理程序.

[2] 朱丹浩, 杨蕾, 王东波. organization recognition model.zip. 基于 RNN 的字模型实体抽取模型.

收稿日期: 2016-08-01
收修改稿日期: 2016-10-26

Recognizing Chinese Organization Names Based on Deep Learning: A Recurrent Network Model

Zhu Danhao[1,2] Yang Lei [3] Wang Dongbo [4]
[1](Library of Jiangsu Police Institute, Nanjing 210031, China)
[2](Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)
[3](Department of High Education, College of Nanjing Traffic Technician, Nanjing 210049, China)
[4](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract: [**Objective**] Chinese organization names are difficult to be recognized by computers due to their complex structures and using of rare words. Successful recognition of these names plays significant roles in information extraction and retrieval, knowledge mining as well as institution research evaluation. [**Methods**] First, we redefined the input and output of organization names based on recurrent neural network method and nature of Chinese words or phrases. Second, we proposed a new model at the word level. **[Results]** Compared to the recurrent network models at the phrase level, the proposed method significantly improved the precision, recall and F value. Among them, the F value increased 1.54%. For organization names with rare words, the F value increased by 11.05%. [**Limitations**] We adopted a greedy strategy to find the local optimal values. A conditional random field method will yield better results from the global perspective. [**Conclusions**] The proposed method, which uses Chinese word level features, is easy to be implemented, and could generate better results than its phrase based counterparts.

Keywords: Organization recognition Recurrent Neural Network Deep learning

ACRL 推出信息素养沙盒框架

大学和研究图书馆协会(Association of College and Research Libraries, ACRL)框架咨询委员会(Framework Advisory Board, FAB)于近日宣布在 sandbox.acrl.org 上推出 ACRL 信息素养沙盒框架。

该沙盒是一个可公开访问的平台和资源库，能帮助图书馆员及其教育合作伙伴在实践和专业发展中发现、共享、收集和使用与 ACRL 高等教育信息素养框架相关的正在进行的工作。该沙盒是一个动态资源，其内容由参与框架的贡献者创建。

ACRL 总裁 Irene M.H. Herold 说：“ACRL 推出了这种创新资源，以支持在各种类型的学术环境中参与该框架的图书馆员的需求。通过提供发现和共享与框架相关的教学和专业开发资源的机会，该沙盒将帮助图书馆员促进信息素养融入学生学习。该沙盒将只限会员使用，所以我们鼓励大家都来参与贡献。”

在这个平台中，游客可以通过搜索符合他们需求的材料进行浏览和贡献，与他人分享自己的材料。当图书馆员发现适用于他们图书馆的案例，或者发现正在研究类似话题的其他人时，该沙盒将促进协作。

有关如何充分利用沙盒的信息，请参阅沙盒帮助中心。

(编译自: http://acrl.ala.org/framework/?p=332)

(本刊讯)